Acceleration of Mobile Commerce using Predictive Retrieval

Amund Tveit and Mihhail Matskin
Norwegian University of Science and Technology

This paper presents predictive retrieval as an approach to improve the experienced end-user bandwidth in mobile commerce. Predictive retrieval is shown to have effect using a simulation of mobile customers accessing a news service.

l

1. INTRODUCTION

With the increasing number of competing mobile services, the process of acquiring new, and retaining existing customers becomes harder. Acquisition of new customers can, to some degree, be done with increased marketing efforts, but marketing has less effect for retaining existing customers, since they allready know the service. Improving the mobile service is likely to be the best way to retain existing customers. The mobile service is improved when the customer gets more "bang for the bucks", e.g. cheaper service, better and fresher content, more support, and the feeling of being a valued customer. S/he also take for granted high reliability and availability of the service, quick response from and fast access to the service. The latter is quite hard due to the relatively low-bandwidth wireless connection from the mobile network provider to the customer with a mobile device. This provider-customer connection is described as the *"last mile problem"* by Bell and Gemmel [1996].

The straightforward way to improve the "last mile" connection is to add bandwidth capacity by replacing the existing wireless network infrastructure, e.g. going from 40-114 kbit/s bandwidth (GPRS) to 384 kbit/s - 2Mbit/s (UMTS). This is unfortunately a very costly and usually unfeasible solution for most providers of mobile services, since it requires major investments in new hardware and software for the provider and that customers have to purchase new mobile devices in order to access the upgraded wireless network.

...

.

In this paper we would like to give answer to the following two questions: Is it possible to improve the "last mile" wireless connection without replacing the existing wireless network infrastructure? If so, is the solution feasible from a business point of view?

2. PROBLEM OUTLINE

Considering a mobile news service, customers can access it using a mobile device with a browser (e.g. Opera). The news content and structure is represented in a hypertext format, e.g. XML-based languages like cHTML, WML or XHTML, where content relations are being represented by unidirectional hyperlinks (e.g. from a menu to an article).

Customers navigate the mobile news service by selecting hyperlinks to news content (e.g. articles or menus) they find of interest, then wait for the selected news content to load, and finally view the content before possibly choose to pursue their navigation. This is shown from a customer perspective in figure 1, where **WAIT** and **READ** is when the customer waits for content to be loaded into the browser and reads the content, respectively.

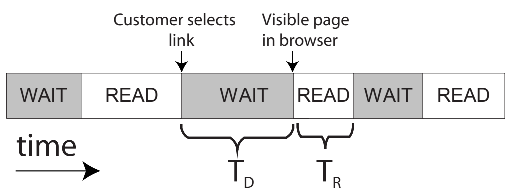

Fig. 1. Customer Navigation Pattern

If the **WAIT** period T_D become too long, the customer may become tired of waiting for the content to be loaded, and decide to go elsewhere, possibly to a competing mobile service. The effect of the relation between download time and consumer attitude has been studied in an e-service context by Rose and Straub [2001], it doesn't conclude that there is such a negative effect after a few delayed download times, but explains that a negative effect may be caused by cumulative, i.e. repeated occurences of delay in download.

2.1 Potential Solution

During the T_R time the customer reads content (**READ**), the "last mile" bandwidth is unused. Hence the average utilized bandwidth is $B_{util} = \frac{\overline{T_D}}{\overline{T_R}+\overline{T_D}}$, e.g. $B_{util} = \frac{1}{3+1} \cdot 114 = 38$ kbit/s if $\overline{T_R} = 3 \cdot \overline{T_D}$ and the "last mile" bandwidth is 114 kbit/s GPRS.

In other words, there is excess bandwidth that can be utilized.

Unused bandwidth between the browser and the mobile network provider gives a possibility for predictive retrieval of data. If the next customer click (i.e. selection of page) can be predicted accurately in advance, the content of the page can be pre-retrieved in the background while the customer is viewing the current page, as shown in figure 2 (right side). Then customer can instantaneously retrieve the page from the browser cache instead of waiting for it to be downloaded. This is significantly much faster than retrieving the page over the wireless network, even if the page would be stored in a local proxy cache which is near network-wise.

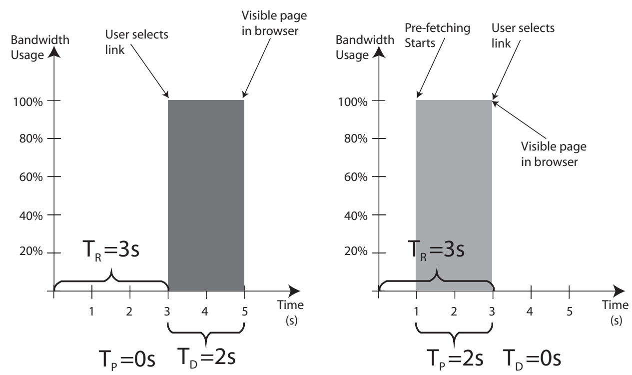

Fig. 2. Browsing without (left) and with (right) optimal predictive retrieval

(T_P is the time spent pre-fetching the predicted content)

Network congestions are usually caused by spikes or prime time use (the case is similar to road traffic rush hours). If the utilized bandwidth was almost evenly distributed without spikes, network congestions would occur much less frequently. In a shared bandwidth environment, e.g. in a company, network congestions would easily occur if every browser uses 100% bandwidth utilization. By using predictive retrieval of content, it is possible to reduce the size of bandwidth spikes by using only a fraction of the available bandwidth when pre-retrieving, but at the same time accelerate browsing, see figure 3 (left).

Unfortunately, it is not realistic to accurately predict which pages and web objects that are going to be selected by the customer in the next click. The effect of inaccurate predictions is shown in figure 3 (right). The result is, as expected, decreased browsing performance as a result of inaccurate prediction.

Personalization is believed to be beneficial for mobile clients (customers) in Dogac and Tumer [2002]. Predictive retrieval is an example of such given that it works, since it has been shown to improve the network performance for mobile customers.

3. BUSINESS MODEL OF PREDICTIVE RETRIEVAL

Algorithms for predictive retrieval in mobile commerce are speculative in the sense that they try to predict *the future* link(s) to page(s) selected by a customer based on *historic* clickstream patterns. This can and will lead to erraneous prefetching of content that the customer is not going to view. The amount of erraneously

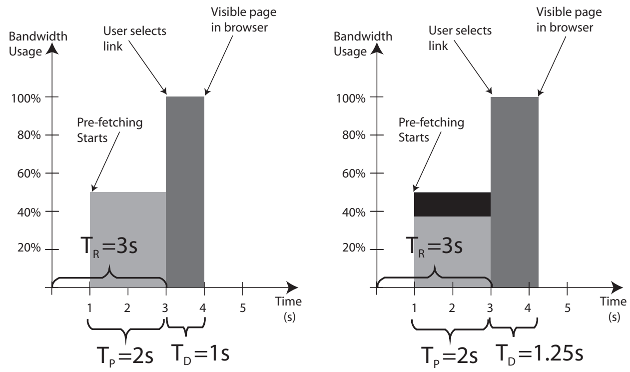

Fig. 3. Browsing with optimal (left) and non-optimal (right) low-bandwidth predictive retrieval

prefetched content can possibly be reduced by adding other prediction approaches, e.g. content-based prediction [Davison 2002], however it isn't likely to be reduced to zero.

In current and future packet-switched mobile networks (e.g. GPRS and UMTS), the billing model is that customers pay per data unit transferred, e.g per MB. If predictive retrieval was introduced with this model it would require the customer to pay *also* when erraneous content is prefetched. Even though the customer could get increased experienced bandwidth, this provides little added value seen from a customer perspective. However, predictive retrieval with time-based billing model commonly used for GSM networks, would not impose additional costs on the customer, since customers don't have to spend more time connected than without predictive retrieval.

On the (wired) Internet it is usually a third party, and not the network provider, who offers content distribution services (e.g. Akamai Inc. for global content replication and caching, and Fireclick Inc. for "last mile" predictive caching). The party who pays for these services is the online web service that wants their content to be faster accessible by their customers, the billing model is usually cost-driven and is a combination of fixed fee per month plus traffic fee (data transferred), or just a traffic fee. The above mentioned model can probably be successfully transferred to a wireless Internet context directly, but it will provide less control over bandwidth utilization within an area, e.g. cell range within parts of a city, than if the predictive retrieval service was hosted by the mobile network provider itself. Then the mobile network provider could use predictive retrieval to improve the utilization of their own wireless networks by turning on predictive retrieval (and the mobile customer's experienced bandwidth) if there is available bandwidth capacity within a cell, and turning it off if the network area is being fully utilized (e.g. in a subway area during rush hours). The first argument for doing this is that the marginal cost of utilizing available bandwidth is approximately zero (analog to filling an empty hotel room or an empty plane sear), and the second argument is that this could be used as a sales argument for mobile network providers towards mobile services and customers to select them instead of another mobile network provider who doesn't provide "accelerated mobile commerce".

4. EXPERIMENT

In order to estimate what the effect of predictive retrieval can be, a simulated model of a news content site has been created, then simulated mobile customers accessing the news content site was created (creating customer access logs), and finally prediction and prefetching was tested on the customer access logs.

4.1 Simulation of News Content Site Structure

It is assumed that the site has an underlying hierarchical structure, with category (i.e menu) and article pages at several levels. There are links both ways between a parent category and its child category or article (as shown with bidirectional black arrows in figure 4).

In addition each subcategory has direct links to *all* parent categories, e.g. *NHL* would also have direct links to *Sports* and the main page (as shown with unidirectional dotted arrows in figure 4).

On category pages there are links directly to hilight articles (top stories), these articles are selected articles from hilight articles of subcategories, e.g. if there is a top story about a *" Surprising Victory!"* in *NHL* it could be that both the main page of the news site and the sports category has direct links to the article (as shown with bidirectional red arrows in figure 4).

The above aspects of a news site has been simulated stochastically, with samples from a normal distribution to determine each nodes fan-out, except for the leaf-nodes, i.e. articles, that only have links to their parents. Hilight articles have been selected by sampling from each sublevel's hilights (if intermediary category) or articles (if bottom category), this has been repeated recursively to propagate and provide hilight articles all the way to the main page at the top level. In order to provide useful data for prefetching, size of pages has been simulated by sampling from a normal distribution.

4.2 Simulation of Browsing Mobile Customers

Need to simulate how many pages a user visits (session length probability), which page they select at each page visit (link selection probability), and how long they stay on each page (page visit length probability).

Session Length Probability Model. In Science Magazine "strong regularities of web surfing" was presented by Huberman et al. [1998], they showed empirically from analysis of several large web logs that the number of pages a user visits is given asymptotically by the inverse gaussian distribution.

$$P(L) = \sqrt{\frac{\lambda}{2\pi L^3}} e^{\left[\frac{-\lambda(L-\mu)^2}{2\mu^2 L}\right]}$$

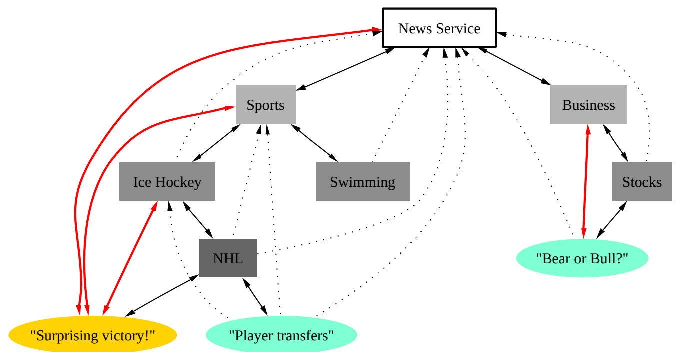

Fig. 4. Example of a News Site Structure

The inverse gaussian distribution has a long right tail, this provides an opportunity for stochastically simulate some customers that visits a large number of pages.

Samples from this distribution has been selected to determine session lengths of simulated users, i.e. how many pages they visit, the underlying assumption for selecting this distribution in the simulation is that mobile internet browsing is similar to web browsing in terms of session length. This is likely to be a simplification, since the number of pages a users visits is not likely to be predetermined, but dependent to how the user responds to each page.

Link Selection Probability Model. The link selection algorithm is based on the multinomial distribution, i.e. selecting links from either the group of hilighted articles, child menus or articles, or parent menus. The probability of selecting a particular group can have uneven probabilities, i.e. the group of highlighted articles is more likely to select than the other, this probability is shown in the legend of resulting plots as $p(hilit = X\%)\,;\, X \in \{50, 60, 80, 95\}$.

Page Visit Length Probability Model. Page visit length T_R is modelled by sampling from a standard gaussian distribution with μ = size of the page. This is assumed based on that the required reading time (in average) for a page is likely to be proportional to the size of the page.

4.3 Predictive Retrieval applied to Simulated Traffic

Need to process the generated customer access logs in order to extract clickstream patterns and customer sessions (lists of visited pages). These clickstream patterns are then used to train prediction algorithms in order to get predictions for what is the most likely next page for a user at any point in the session (i.e. given an ordered set of visited pages within that session).

In order to utilize the clickstream data at its fullest in terms of prediction and prefetching accuracy, a n-fold cross validation method is selected (Cohen [1995]). The method creates n=10 independent estimates of prediction and prefetch accuracy, where each estimate is based on training prediction algorithms with 90% of the clickstream data and and testing on the last 10% of the data.

The selected algorithms for prediction are 1-step Markov (selected for simplicity, [Taylor and Karlin 1993]) and Support Vector Machines (selected for anticipated high accuracy, using libSVM [Chang and Lin 2001]). Prediction in this context is equivalent to a classification problem: given a historic clickstream path for a user, classify which next page (or next pages sorted after probability) the user visits.

Measurements are based on recall and precision from the information retrieval field, where *recall is the fraction of relevant data that has been retrieved* and *precision is fraction of retrieved data that was relevant* [Baeza-Yates and Ribeiro-Neto 1999]

5. RESULTS

The experiment was performed with 1 stochastically generated news site accessed by 1000 simulated mobile customers. Experimental details including source code (in Python), configuration files (with statistical parameter settings), statistical analysis (with R/RPy), output logs and plots (prediction accuracy with Markov and SVM) `http://gamemining.net/msim/`.

Does predictive retrieval have positive effect on performance? If there is significant evidence that $recall > 0$ and $precision > 0$ so yes, otherwise no. By performing a Wilcox hypothesis test (non-parametric, assumes unknown variance and nothing about the underlying distributions) of both recall and precision with 95% confidence interval gives: $recall \in [7.99\%, 100.00\%]$ and $recall \in [20.00\%, 100.00\%]$, both well above 0. Yes, predictive retrieval has improved performance.

6. CONCLUSION

We believe to have shown that answered is yes to the two initial questions. It is possible to improve the "last mile" wireless connection with predictive retrieval (as shown in a simulated experiment), and it is also economically feasible (as discussed in business models). However, it is yet to be shown that this can be transferred with success from a simulated a real-life mobile news service. The contribution of this paper has been threefold, first we have shown using stochastic simulation that predictive retrieval can accelerate mobile services, second we have developed and made public simulation code for the experiment, and third we have discussed

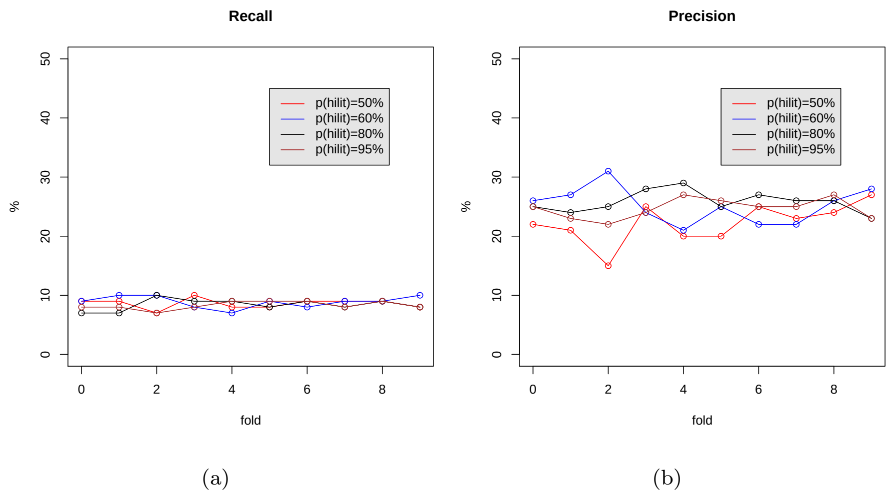

Fig. 5. Precision and Recall Measurements (1-step Markov)

business issues related to predictive retrieval in a mobile service context.

Opportunities for future work include continuing to improve the realism of the mobile news service/customer simulator and run more extensive experiments with data generated from it, test other prediction/classification methods, and finally implement a prototype of the system. It is also of interest to investigate whether predictive retrieval approaches works in other and more complex domains than the mobile web, e.g. in mobile multiplayer games [Tveit and Tveit 2002].

ACKNOWLEDGMENTS

This work is partially supported by the Norwegian Research Council.

REFERENCES

Baeza-Yates, R. and Ribeiro-Neto, B. 1999. *Modern Information Retrieval*, 1st ed. ACM Press, Chapter 3 - Retrieval Evaluation.

Bell, G. and Gemmel, J. 1996. On-ramp prospects for the information superhighway dream. *Communications of the ACM 39,* 7 (June), 55–61.

Chang, C.-C. and Lin, C.-J. 2001. Training nu-support vector classifiers: Theory and algorithms. *Neural Computation 13,* 9, 2119–2147.

Cohen, P. R. 1995. *Empirical Methods for Artificial Intelligence*. MIT Press, Chapter 6.10 - Cross-Validation: An Efficient Training and Testing Procedure, 216–218.

Davison, B. D. 2002. Predicting web actions from html content. In *The Thirteenth ACM Conference on Hypertext and Hypermedia*. 159–168.

Dogac, A. and Tumer, A. 2002. Issues in Mobile Electronic Commerce. *Journal of Database Management 13,* 1 (January), 37–43.

Huberman, B. A., Pirolli, P. L. T., Pitkow, J. E., and Lukose, R. M. 1998. Strong Regularities in World Wide Web Surfing. *Science 280,* 5360, 95–97.

Rose, G. M. and Straub, D. W. 2001. The effect of download time on consumer attitude toward the e-service retailer. *e-Service Journal 1,* 1, 55–76.

Taylor, H. M. and Karlin, S. 1993. *An Introduction to Stochastic Modelling*, Revised ed. Academic Press, Chapter 3 - Markov Chains.

Tveit, A. and Tveit, G. B. 2002. Game Usage Mining: Information Gathering for Knowledge Discovery in Massive Multiplayer Games. In *Proceedings of the 3rd International Conference on Internet Computing*, H. R. Arabnia and Y. Mun, Eds. Vol. 3. CSREA Press, 636–642.